UNIVERSITY OF TECHNOLOGY SYDNEY
CAI, SoS, FEIT

LEARNING AND REPRESENTING ATTRIBUTED GRAPHS

by

Ruiqi Hu

A Thesis Submitted
in Partial Fulfillment of the
Requirements for the Degree

Doctor of Philosophy

Dissertation Directed by
Dr. Guodong Long, Dr. Shirui Pan, and Prof. Chengqi Zhang

Sydney, Australia

2018

Certificate of Authorship/Originality

I certify that the work in this thesis has not been previously submitted for a degree nor has it been submitted as a part of the requirements for other degree except as fully acknowledged within the text.

I also certify that this thesis has been written by me. Any help that I have received in my research and in the preparation of the thesis itself has been fully acknowledged. In addition, I certify that all information sources and literature used are quoted in the thesis.

This research is supported by the Australian Government Research Training Program.

Ruiqi Hu

Production Note:
Signature removed
prior to publication.

ABSTRACT

LEARNING AND REPRESENTING ATTRIBUTED GRAPHS

by

Ruiqi Hu

Information graphs are ubiquitous in many areas, such as medicine, social media and academic engines, and each node in the graph comes with various attributes. For example, in a academic citation graph, we can take each paper a node, then the author(s) and title of each paper can be extracted as the attributes of the node. Moreover, papers, authors as well as venues can be taken as different sources of nodes in one information graph. By doing so, we have got a heterogeneous information graph with more than one sources of nodes, attributes and links.

To implement these applications, such as identifying protein residues and social media marketing, graph representation of homogeneous information graphs has been widely researched and employed. This research, aims to embed and represent homogeneous nodes with low-dimensional and unified vectors, while preserving the contextual information between nodes, and, as a result, classical machine learning methods can be directly applied.

However, existing graph embedding algorithms are facing five major challenges: 1.the graph representation learning and node classification in graphs are separated into two steps, which may result in sub-optimal results because the node representation may not fit the classification model well; 2. existing ones are mostly shallow methods that can only capture the linear and simple relationships in the data; 3.ignoring the data distribution of the latent codes from the graphs, which often results in inferior embedding in real-world graph data; 4. unable to handle the heterogeneous and multi-relational information graph which is the major form that graph data existed in the real-world; and 5. unable to effectively discover functional groups

and understand the roles of detected groups.

To face the aforementioned challenges, the main research objective of the thesis is to study that how to more effectively embed the nodes of a graph into a compact space for the tasks which are most related to the real-world applications.

The main research objective has been studied from four coherently linked perspectives: (1) How to unify the traditional two-step embedding work-flow into one smooth embedding procedure to avoid the inconsistency between the embedding architecture and classifier; (2) How to learn a universal embedding for all sources of nodes in a graph, so one single embedding can be used to represent the entire heterogeneous information graph; (3) How to smoothly regularize the embedding with a certain distribution during the learning procedure for a more robust embedding;(4) How to automatically generate a human-understandable explanation of each cluster of nodes in the graph and applied the algorithm in the real business world.

Specifically, this thesis aims to tackle aforementioned challenges by conducting studies of graph ladder network to unifies both representation and classifier model learning into one framework; developing universal graph representation to represent different types of nodes in heterogeneous information graph in a continuous and common vector space; introducing generative adversarial scheme into graph domain to encode the topological structure and node content in a graph to a compact representation, on which a decoder is trained to reconstruct the graph structure under an adversarial training scheme and carrying out co-clustering on enterprise information graph for functional group discovery and understanding. All works in this thesis are validated with related tasks like graph classification, graph clustering, graph visualization and link prediction respectively.

Acknowledgements

Firstly, I would like to express my sincere gratitude to my superviors Dr. Guodong Long, Dr. Shirui Pan, and Prof. Chengqi Zhang for their continued support and guidance in my Ph.D. study and related projects, specifically for their patience, motivation, and immense knowledge. Their guidance helped me in the entire research and writing of this thesis. I could not have imagined having better supervisors or mentors for my Ph.D study.

A very special gratitude goes to Prof. Mingan Choct, who has provided life-changing support and opportunities. Words cannot express my emotions in this regard.

Thanks to all those who shared my sleepless nights in the laboratory. It was great spending time in the laboratory with all of you.

I am also grateful to all the following people, who have helped and supported me along the way: Dr. Jing Jiang, Prof. Li Xue, Dr. Tony W.T. Chan, Mr. David Xue, Dr. Xueping Peng, and the list goes on.

And finally, last but by no means least, I would like to thank my family: my father Mr. Wenming Hu, who always encourages me to try harder in the business world; my mother Prof. Shuqing Yang, who strongly believes that deep learning is not science; and my fiancée Qian Zhang; without her, it is impossible for me to just focus on what I am really interested in.

Thanks for all your encouragement!

Ruiqi Hu

Sydney, Australia, 2018.

List of Publications

Journal Papers

J-1. [TCYB] **Ruiqi Hu**, Shirui Pan, Sai-fu Fung, Guodong Long, Jing Jiang, and Chengqi Zhang. "Learning Graph Embedding with Adversarial Training Methods", IEEE Transactions on Cybernetics. (Under review) (ERA Rank **A*** journal).

J-2. [PR] **Ruiqi Hu**, Shirui Pan, Sai-fu Fung, Guodong Long. "Clustering Social Audiences in Business Information Networks", Pattern Recognition. (Under review) (ERA Rank **A*** journal).

J-3. [TNNLS] Chun Wang, **Ruiqi Hu**, Shirui Pan, Guodong Long and Jing Jiang. "Unsupervised Deep Neighbor-aware Embedding for Graph Clustering" IEEE Transactions on Neural Networks and Learning Systems. (Under review) (the first and second author contribute equally to this work) (ERA Rank **A*** journal)

Conference Papers

C-1. [IJCAI-19] **Ruiqi Hu**, Shirui Pan, , Guodong Long, Jing Jiang, Lina Yao and Chengqi Zhang. "Going Deep: Graph Convolutional U-Shape Networks for Semi-supervised Node Classification" 2019 International Joint Conference on Artificial Intelligence. (Under review) (CORE Rank **A*** conference)

C-2. [IJCAI-19] Chun Wang, Shirui Pan, **Ruiqi Hu**, Guodong Long, Jing Jiang and Chengqi Zhang. "DAEGC: Unsupervised Deep Attentional Embedding for Graph Clustering" 2019 International Joint Conference on Artificial Intelligence. (Under review) (CORE Rank **A*** conference)

C-3. [IJCAI-18] Shirui Pan, **Ruiqi Hu**, Guodong Long, Jing Jiang, Lina Yao, and Chengqi Zhang. "Adversarially Regularized Graph Autoencoder." International Joint Conference on Artificial Intelligence, 2018 (accepted on 17 April 2018) (Shirui Pan and Ruiqi Hu contribute equally to this work) (ERA Rank **A***)

C-4. [CIKM-17] **Ruiqi Hu**, Shirui Pan, Jing Jiang, and Guodong Long. "Graph Ladder Networks for Network Classification." In Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, pp. 2103-2106. ACM, 2017 (ERA Rank A).

C-5. [IJCNN-16] **Ruiqi Hu**, Shirui Pan, Guodong Long, Xingquan Zhu, Jing Jiang, and Chengqi Zhang. "Co-clustering enterprise social networks." In Neural Networks (IJCNN), 2016 International Joint Conference on, pp. 107-114. IEEE, 2016 (ERA Rank A).

C-6. [IJCNN-17] **Ruiqi Hu**, Celina Ping Yu, Sai-Fu Fung, Shirui Pan, Haishuai Wang, and Guodong Long. "Universal network representation for heterogeneous information networks." In Neural Networks (IJCNN), 2017 International Joint Conference on, pp. 388-395. IEEE, 2017 (ERA Rank A).

C-7. [IJCNN-19] Di Wu, Ruiqi Hu, Yu Zheng, Jing Jiang and Michael Blumenstein. "Feature-Dependent Graph Convolutional Autoencoders with Adversarial Training Methods" 2017 International Joint Conference on, (accepted on 08 March 2019) (Di Wu and Ruiqi Hu contribute equally to this work) (CORE Rank A conference)

Contents

List of Figures

Abbreviation

EM - Expectation Maximization

GCN - Graph Convolutional Network

NMF - Non-negative Matrix Factorization

CNN - Convolutional Neural Network

MLP - Multi-layer Perceptron

MHIG - Multi-relational Heterogeneous Information Graph

SGD - Stochastic Gradient Descent

SAC - Social Audience Cluster

BIG - Business Information Graph

K-NN - K-Nearest Neighbors

PPMI - Positive Point-wise Mutual Information

Nomenclature and Notation

Capital letters denote matrices.

Lower-case alphabets denote column vectors.

$(.)^T$ denotes the transpose operation.

I_n is the identity matrix of dimension $n \times n$.

0_n is the zero matrix of dimension $n \times n$.

$\mathbb{R}$, $\mathbb{R}^+$ denote the field of real numbers, and the set of positive reals, respectively.